

The FastHCS Algorithm for Robust PCA

Eric Schmitt and Kaveh Vakili

the date of receipt and acceptance should be inserted later

Abstract Principal component analysis (PCA) is widely used to analyze high-dimensional data, but it is very sensitive to outliers. Robust PCA methods seek fits that are unaffected by the outliers and can therefore be trusted to reveal them. FastHCS (High-dimensional Congruent Subsets) is a robust PCA algorithm suitable for high-dimensional applications, including cases where the number of variables exceeds the number of observations. After detailing the FastHCS algorithm, we carry out an extensive simulation study and three real data applications, the results of which show that FastHCS is systematically more robust to outliers than state-of-the-art methods.

Keywords: High-dimensional data, outlier detection, computational statistics, exploratory data analysis

1 Introduction

Principal component analysis (PCA) is widely used to explore high-dimensional data. It centers and rotates the original p -dimensional measurements to construct a small number q of new orthonormal variables, called *principal components*, that account for most of the variation in the data. However, classical PCA is very sensitive to outliers. Outliers are observations that are inconsistent with the multivariate pattern of the majority of the data. If left unchecked, they influence the estimated parameters by disproportionately pulling the fit towards themselves. In this way, outliers obscure the

main relationships in the data and their true outlyingness. In practice, we want to find the outliers to bound their influence on the fit and to study as objects of interest in their own right. For these reasons, we need robust PCA methods that meet the following criteria: (1) Like classical PCA, a robust PCA method should handle cases where the number of variables exceeds the number of observations, (2) and it should be shift and rotation equivariant, meaning that if the data are shifted or rotated the estimated parameters should transform accordingly. (3) It should be computable for high-dimensional data. (4) It should accurately describe the multivariate pattern of the majority of the observations, even when the data is heavily contaminated by outliers. (5) It should have a high breakdown point; a measure an estimator's robustness to outliers in the data. (6) It should be insensitive to the dimensionality of the data.

Criteria (1)-(3) are natural for any PCA method. Criteria (4)-(5) relate to robustness. Criterion (6) is related to both concerns. We find that state-of-the-art robust PCA algorithms have most of these properties, but that, surprisingly, many instances can be found where they fail to satisfy Criterion (4). In this paper, we introduce a robust PCA algorithm, FastHCS, to meet these criteria (HCS for high-dimensional congruent subset). In the next section we outline FastHCS. Then, in Sections (3) and (4) we compare it to several state-of-the-art methods on simulated data and three real data applications which show that in many settings only FastHCS can be relied upon to provide a robust PCA solution.

E. Schmitt
 Protix
 Industriestraat 3
 5107 NC Dongen Tel.: +31162782501
 E-mail: eric.schmitt@protix.eu

2 FastHCS

Given an $n \times p$ data matrix $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^n$ and for a fixed $2 \leq q < \min(p, n)$, the FastHCS algorithm searches for a subset of size at least $h = \lceil (n + q + 1)/2 \rceil$ free of outliers (this is the minimal value of h such that there are at least $(q + 1)$ clean observations in each candidate subset).

If $p > n$, FastHCS computes the mean-centered data matrix $\tilde{\mathbf{X}} = \mathbf{Y} - \mathbf{1}_n^\top (\text{ave}_{i=1}^n \mathbf{y}_i)$, and performs the kernel eigenvalue decomposition of $\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top = \mathbf{U}\mathbf{L}\mathbf{U}^\top$ where \mathbf{U} is an $n \times r$ matrix, \mathbf{L} is an $r \times r$ matrix, $r := \text{rank}(\mathbf{X})$ and \mathbf{L} is a diagonal matrix with the eigenvalues on the diagonal. Then, FastHCS works with the $n \times r$ matrix $\mathbf{X} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top \mathbf{U}(\mathbf{L})^{-1/2}$. The transformation from \mathbf{Y} to \mathbf{X} causes no loss of information or robustness since we retain all of the components corresponding to non-zero eigenvalues. However, this transformation reduces the computational cost of the subsequent steps of the algorithm. At the end of the algorithm, FastHCS reverses these transformations so that the returned parameter estimates are consistent with conventional PCA. When $n \leq p$, we simply set $\mathbf{X} = \mathbf{Y}$.

2.1 The I -index h -subset

The I -index is a subset selection criterion first introduced in Vakili and Schmitt (2014) where it is used to identify an outlier free subset to serve as the basis of the robust PCS location and scatter estimator. The I -index was designed to be insensitive to the configuration of the outliers and consequently, as we show in that article, the fit found by FastPCS is nearly unaffected by the presence of outliers in the data (we refer to this as quantitative robustness). In (Schmitt et al., 2014) we further show that the PCS estimates also have the maximum possible breakdown point (we refer to this as qualitative robustness). Robust location and scatter estimation are also important for PCA. In the PCS context, the I -index is applied to the observations in their original dimensionality, and one approach to achieving robust PCA would be to use the robust PCS covariance estimate as a starting point for PCA. However, this approach does not satisfy Criterion (3) for robust PCA since it is not possible to perform PCS when the number of dimensions is greater than the number of observations. This subsection describes how the I -index can be extended to the PCA context by applying it to projections of the data on to subspaces.

To begin, FastHCS draws M random subsets of size $(q + 1)$ from \mathbf{X} without replacement, where M is given

by:

$$M = \left\lceil \frac{\log(0.01)}{\log(1 - (e/n)^{q+1})} \right\rceil, \quad (1)$$

and where $h \leq e < n$ is an integer specifying the number of uncontaminated observations, so that the probability of getting at least one uncontaminated starting subset is at least 99% (Stahel, 1981). By default we set $e = h$. However, if the user is sure that the contamination rate of the sample is lower than $(n - h)/n$, we offer the possibility (as in Maronna and Yohai (1995)) of using this information to reduce the computational cost of running FastHCS. Denote these $(q + 1)$ -subsets as $\{H_0^m\}_{m=1}^M$. The SVD decomposition of the observations indexed by H_0^m is:

$$\text{svd}_{i \in H_0^m}((\mathbf{x}_i - \mathbf{t}_0^m)/\sqrt{q}) = \mathbf{U}_0^m(\mathbf{L}_0^m)^{1/2}(\mathbf{P}_0^m),$$

where $\mathbf{t}_0^m = \text{ave}_{i \in H_0^m} \mathbf{x}_i$ is the estimated center, \mathbf{L}_0^m is a diagonal matrix for which the non-zero elements $(\mathbf{L}_0^m)_j$ $j = 1, \dots, q$ are the descending eigenvalues of the PCA model fitted to $\{\mathbf{x}_i : i \in H_0^m\}$, and the eigenvectors $\mathbf{P}_{0,q}^m$ are the first q loadings of this model. Next, we compute the score matrix \mathbf{S}_0^m with n rows $\mathbf{s}_{0,i}^m$:

$$\mathbf{s}_{0,i}^m = (\mathbf{x}_i - \mathbf{t}_0^m)\mathbf{P}_{0,q}^m, \quad 1 \leq i \leq n$$

which is the projection of the re-centered rows of \mathbf{X} on to the subspace spanned by the first q loadings of $\{\mathbf{x}_i : i \in H_0^m\}$. To measure the outlyingness of an $\mathbf{s}_{0,i}^m$ to the members of $\{\mathbf{s}_{0,i}^m : i \in H_0^m\}$, we will use its squared orthogonal distance to \mathbf{a}_k^m , the direction normal to the hyperplane through q members of $\{\mathbf{s}_{0,i}^m : i \in H_0^m\}$ drawn at random:

$$d_i^2(\mathbf{a}_k^m, \mathbf{S}_0^m) = ((\mathbf{s}_{0,i}^m)^\top \mathbf{a}_k^m - 1)^2 / \|\mathbf{a}_k^m\|^2,$$

and, to remove the dependence of this measure on the direction \mathbf{a}_k^m , we average it over K such directions $\{\mathbf{a}_k^m\}_{k=1}^K$:

$$D_i(H_0^m) = \frac{K}{\text{ave}_{k=1}^K} \frac{d_i^2(\mathbf{a}_k^m, \mathbf{S}_0^m)}{\text{ave}_{i \in H_0^m} d_i^2(\mathbf{a}_k^m, \mathbf{S}_0^m)}, \quad 1 \leq i \leq n. \quad (2)$$

(In Remark 1 below we discuss how we set the value of the parameter K). The denominator in Equation (2) normalizes these distances across the directions \mathbf{a}_k^m .

We can now describe the computation of the first step of FastHCS. For a given a $(q + 1)$ -subset H_0^m of $\{1 : n\}$ and its corresponding matrix \mathbf{S}_0^m , Algorithm 1 returns an h -subset H^m of indexes of $\{1 : n\}$ using an iterative procedure we call *growing steps*. In each step w , H_w^m is updated and contains the indexes of the ω_w observations with smallest values of $D_i(H_{w-1}^m)$. The value of ω_w itself increases incrementally from

$\lceil (n - q - 1)/(2W) \rceil + q + 1$ to h in W steps. These steps do not have a convergence criterion, so the number of iterations W must be set in advance (In Remark 1 below we discuss how we set the value of the parameter W).

Algorithm 1: growingStep(H_0^m, \mathbf{X}, q)

for $w = 1$ to W do:

$$D_i(H_{w-1}^m) \leftarrow \frac{K}{\text{ave}_{k=1}^K} \frac{d_i^2(\mathbf{a}_k^m, \mathbf{S}_0^m)}{\text{ave}_{i \in H_{w-1}^m} d_i^2(\mathbf{a}_k^m, \mathbf{S}_0^m)}, \quad 1 \leq i \leq n$$

set $\omega_w \leftarrow \lceil (n - q - 1)w/(2W) \rceil + q + 1$

set $H_w^m \leftarrow \{i : D_i(H_{w-1}^m) \leq D_{(\omega_w)}(H_{w-1}^m)\}$

end for

$H^m \leftarrow H_W^m$

After growing M candidate H^m 's, FastHCS evaluates each using a criterion we call the I -index, and fits a robust PCA model to the H^m having smallest value of the I -index. For a given h -subset H^m and direction \mathbf{a}_k^m , we define a subset H_k^m that is optimal with respect to \mathbf{a}_k^m in the sense that it indexes the h observations with the smallest values of $d_i^2(\mathbf{a}_k^m, \mathbf{S}_0^m)$. More precisely, denoting $d_{(h)}$ the h^{th} order statistic of a vector \mathbf{d} , we have:

$$H_k^m = \{i : d_i^2(\mathbf{a}_k^m, \mathbf{S}_0^m) \leq d_{(h)}^2(\mathbf{a}_k^m, \mathbf{S}_0^m)\}.$$

Then, we define the I -index of an H^m along \mathbf{a}_k^m as

$$I(H^m, \mathbf{S}_0^m, \mathbf{a}_k^m) = \log \left(\frac{\text{ave}_{i \in H^m} d_i^2(\mathbf{a}_k^m, \mathbf{S}_0^m)}{\text{ave}_{i \in H_k^m} d_i^2(\mathbf{a}_k^m, \mathbf{S}_0^m)} \right), \quad (3)$$

with the convention that $\log(0/0) := 0$. The measure $I(H^m, \mathbf{S}_0^m, \mathbf{a}_k^m)$ is always positive and increases the fewer members H^m shares with H_k^m along the direction \mathbf{a}_k^m . This is because, for a given direction \mathbf{a}_k^m , the members of H_k^m not in H^m will decrease the denominator in Equation (3) without affecting the numerator, increasing the overall ratio. As in the growing steps, we remove the dependence of Equation (3) on the directions \mathbf{a}_k^m by considering the average over K directions:

$$I(H^m, \mathbf{S}_0^m) = \frac{K}{\text{ave}_{k=1}^K} I(H^m, \mathbf{S}_0^m, \mathbf{a}_k^m). \quad (4)$$

Finally, FastHCS selects as H^I the candidate h -subset H^m with lowest I -index.

Given H^I , we denote the PCA parameters corresponding to H^I as $(\mathbf{t}^I, \mathbf{L}_q^I, \mathbf{P}_q^I)$ and obtain them as follows:

$$\text{svd} \left((\mathbf{y}_i - \mathbf{t}^I) / \sqrt{h-1} \right)_{i \in H^I} = \mathbf{U}^I (\mathbf{L}^I)^{1/2} (\mathbf{P}^I)^\top,$$

where $\mathbf{t}^I = \text{ave}_{i \in H^I} \mathbf{y}_i$. FastHCS computes these parameters on the full space of the data set, \mathbf{Y} , rather than on the space of \mathbf{S}_0^I , to increase their accuracy. Algorithm 2 summarizes the I -index step of FastHCS.

Algorithm 2: IStep(\mathbf{X}, q)

for $m = 1$ to M do:

$H_0^m \leftarrow \{\text{random } (q+1)\text{-subset from } 1 : n\}$

$H^m \leftarrow \text{growingStep}(H_0^m, \mathbf{X}, q)$

$I(H^m, \mathbf{S}_0^m) \leftarrow \frac{K}{\text{ave}_{k=1}^K} I(H^m, \mathbf{S}_0^m, \mathbf{a}_k^m)$

end for

$H^I \leftarrow \underset{H^1, \dots, H^M}{\text{argmin}} I(H^m, \mathbf{S}_0^m)$

return $(\mathbf{t}^I, \mathbf{L}_q^I, \mathbf{P}_q^I)$

Remark 1 Through experiments, we find that increasing K above 25 or W above 5 does not noticeably improve the performance of the algorithm (though it increases its computational cost), so we set these parameters to those values. Those experiments were carried on the outlier configurations discussed in Sections 3 and 4 as well as additional configurations enabled by the simulation suite provided with the Online Resources (Section 4). Because such experiments cannot cover all possible configurations of outliers, we focused on those configurations singled out as most challenging in the literature on robust PCA.

Remark 2 Exact fit: When the h members of a subset H' lie on a subspace $\boldsymbol{\Pi}_r \in \mathbb{R}^r$ with $1 < r \leq q$ the numerator and denominator of $I(H', \mathbf{S}_0', \mathbf{a}_k')$ will be the same for any direction \mathbf{a}_k' through members of H' (Schmitt et al., 2014) so that $I(H', \mathbf{S}_0') = 0$. Then, $H^I = H'$ and \mathbf{P}_r^I will correspond with $\boldsymbol{\Pi}_r$. In such situations, FastHCS will return the index of the members of $\{i : ((\mathbf{x}_i - \mathbf{t}^I) \mathbf{P}_r^*)^2 = 0\}$. This behavior is called exact fit (Maronna et al., 2006).

Remark 3 Breakdown point: The (finite sample) breakdown point of an estimator referred to in Criterion (5) is the smallest proportion of observations that need to be replaced by arbitrary value to drive the estimates to arbitrary values (Donoho, 1982). Naturally, a higher breakdown point is better, and the maximal breakdown point achievable in the PCA context is essentially fifty percent.

Both the growing step and the I -index use distances computed on subspaces to derive a measure of outlyingness. Since they are restricted to this view of the data, they are vulnerable to outliers that appear consistent with the majority on a subspace, but are outlying with respect to it (Appendix A details the specific configurations of outliers giving rise to this issue). Therefore, fits

based on H^I alone will not have maximum breakdown and the procedure presented above must be combined with a second, computational expedient, ancillary procedure to ensure that the final FastHCS estimates do.

2.2 The Projection Pursuit h -subset

Although experiments, such as those in Sections 3 and 4, show that the I -index rarely selects contaminated subsets, it is vulnerable to specific configurations of outliers (see Appendix A). To guard against these, FastHCS uses a robust projection-pursuit (PP) approach to identify a second subset of observations, H^{PP} . The PP approach proceeds by assigning an outlyingness score to each observation:

$$d_i^{PP}(\mathbf{Y}) = \max_{\mathbf{v} \in B} \frac{|\mathbf{y}_i \mathbf{v} - \text{med}(\mathbf{y}_j \mathbf{v})|}{\text{mad}(\mathbf{y}_j \mathbf{v})} \quad (5)$$

where B contains 1000 directions \mathbf{v} , each given by two data points drawn randomly from the sample, $\text{med}(\mathbf{y}_j \mathbf{v})$ is the median of $\{\mathbf{y}_j \mathbf{v}, j = 1, \dots, n\}$ and $\text{mad}(\mathbf{y}_j \mathbf{v}) = \text{med}|\mathbf{y}_j \mathbf{v} - \text{med}(\mathbf{y}_l \mathbf{v})|$. The PP method is orthogonally invariant and computationally expedient. A version of the PP algorithm is used as an initial step in ROBPCA (Hubert et al. , 2005; Debruyne and Hubert , 2009), a popular robust PCA algorithm.

2.3 Selecting the final PCA model

Consider the subset $H^\bullet := H^I \cap H^{PP}$. Because $h \geq \lceil n/2 \rceil + 1$, it holds that $|H^\bullet| \geq q$ and H^\bullet is free of outliers whenever either one of H^I or H^{PP} is. We propose to exploit this fact to select between the I -index and PP-based models. Denote $H^- = H^{PP} \setminus H^I$ and

$$D(\mathbf{Y}, H^I, H^{PP}) = \max_{j=1}^q \log \frac{\text{ave}_{i \in H^I} ((\mathbf{y}_i - \mathbf{t}^I) \mathbf{P}_j^I)^2}{\text{var}_{i \in H^\bullet} (\mathbf{y}_i \mathbf{P}_j^I)} - \max_{j=1}^q \log \frac{\text{ave}_{i \in H^\bullet} ((\mathbf{y}_i - \mathbf{t}^{PP}) \mathbf{P}_j^{PP})^2}{\text{var}_{i \in H^-} (\mathbf{y}_i \mathbf{P}_j^{PP})}, \quad (6)$$

with the assumption that $\log(0/0) = 0$. When $D(\mathbf{Y}, H^I, H^{PP}) > 0$ (or if $\max_{j=1}^q \text{var}_{i \in H^-} (\mathbf{y}_i \mathbf{P}_j^{PP}) = 0$) the final FastHCS parameters $(\mathbf{t}^*, \mathbf{L}_q^*, \mathbf{P}_q^*)$ will be equal to $(\mathbf{t}^{PP}, \mathbf{L}_q^{PP}, \mathbf{P}_q^{PP})$ and the final FastHCS subset H^* is set as H^{PP} . As we show in Appendix B, this selection rule ensures that the FastHCS fit has a high breakdown point. Our approach is similar to the ROBPCA algorithm which also selects from among two candidate subsets in the final stage of the algorithm. ROBPCA selects the subset whose eigenvalues have the smallest product. However, depending on the configuration of

the outliers and the rate of contamination, it is possible for a contaminated subset to have smaller eigenvalues than an uncontaminated one (Schmitt et al. , 2014), and so to end up being selected by the criterion used in ROBPCA. In contrast, the selection criterion we propose controls (through the denominators in Equation (6)) for the relative scatter of the two subsets and so it is not biased towards subsets containing many concentrated outliers. Naturally, criterion (6) is designed to favor the I -index based model whenever doing so does not cause breakdown.

2.4 Outlyingness to the PCA model

Two concepts of distance are used to assess the outlyingness of an observation with respect to a PCA model, and cut-off values for both of these can be used to classify outliers (Hubert et al. , 2005). The first is the orthogonal distance (OD) of the observation to the PCA model space:

$$\text{OD}_i(\mathbf{t}, \mathbf{P}_q) = \|\mathbf{x}_i - \mathbf{t} - (\mathbf{x}_i - \mathbf{t}) \mathbf{P}_q (\mathbf{P}_q)^\top\| \quad (7)$$

Assuming multivariate normality of the observations on which the PCA model is fitted, a cut-off can be obtained for the OD statistics using the Wilson-Hilferty transformation of the ODs into approximately normally distributed random variables:

$$c_e(\mathbf{t}, \mathbf{P}_q, H) = \left(\text{ave}_{i \in H} \text{OD}_i^{2/3}(\mathbf{t}, \mathbf{P}_q) + \Phi_{0.975}^{-1} \sqrt{\frac{\text{var}_{i \in H} \text{OD}_i^{2/3}(\mathbf{t}, \mathbf{P}_q)}{\chi_{e/n, 1}^2}} \right)^{3/2} \quad (8)$$

where $\chi_{e/n, 1}^2$ is the e/n quantile of the χ^2 distribution with one degree of freedom, and H indexes a subset of observations. The second measure of outlyingness is the score distance (SD) of the observation on the PCA model space:

$$\text{SD}_i(\mathbf{t}, \mathbf{L}_q, \mathbf{P}_q) = \sqrt{\frac{((\mathbf{x}_i - \mathbf{t}) \mathbf{P}_q) (\mathbf{L}_q)^{-1}}{((\mathbf{x}_i - \mathbf{t}) \mathbf{P}_q)^\top}}. \quad (9)$$

A 97.5% cut-off for the SD statistics can be obtained using a $\sqrt{\chi_{0.975, q}^2}$ distribution.

In inferential applications, PCA theory typically assumes multivariate normality, though ellipticity is sufficient for many of the hypotheses of interest to PCA-based inference, see (Jensen , 1986) and (Jolliffe , 2002, pages 49,55,394). In any case, robust PCA performs inference with a model fitted on the non-outlying observations, so the distributional assumption pertains to only this subset of the data. Conversely, no assumptions are made on the distribution(s) of the outliers.

2.5 Computational considerations

The computational complexity of FastHCS is determined by the I -index and PP subset selection components. The complexity of the PP-based approach is $\mathcal{O}(qnp)$. This is dominated by the time complexity of the I -index, which scales as $\mathcal{O}(q^3 + nq)$ for each starting $(q + 1)$ -subset. Except when q and n are small (smaller than about 5 and 2000 in our tests) FastHCS is not the quickest of the robust PCA methods we considered (in general, we find that PcaL is). The 'Fast' qualification in this context ("FastHCS") is used to distinguish the algorithm based on random sub-sampling from the naïve one based on exhaustive enumeration of all possible starting points, the latter being usually not computable in practice. The computing time of FastHCS grows similarly in n to other methods we discuss in this paper, while it is the most sensitive to increases in q , with computation times being comparable until around $q = 12$. For higher q , FastHCS is the slowest overall. In practice, FastHCS becomes impractical for values of q much larger than 25. Nevertheless, the overall time complexity of FastHCS grows with q , instead of p , making it a suitable candidate for high-dimensional applications, and satisfying Criterion (3) for a robust PCA method. Moreover, FastHCS belongs to the class of so called 'embarrassingly parallel' algorithms, i.e. its time complexity scales as the inverse of the number of processors, meaning it is well suited to benefit from modern computing environments. To enhance user experience, we implemented FastHCS in C++ and wrapped it in an portable, open source R package (R Core Team, 2012) distributed through CRAN (package **FastHCS**)

3 Simulation Study

In this section we evaluate the behavior of FastHCS against three other robust PCA methods: the ROBPCA (Hubert et al., 2005), PcaPP (Croux and Ruiz-Gazen, 2005) and PcaL (Locantore et al., 1999) algorithms. Although other methods for high-dimensional outlier detection exist, these are particularly comparable with FastHCS: all three are PCA algorithms, satisfying Criteria (1)-(3) of a robust PCA method. ROBPCA, PcaPP and PcaL were computed using the R (R Core Team, 2012) package **rrcov** (Todorov and Filzmoser, 2009) with default settings except for the robustness parameter **alpha** for ROBPCA which we set to 0.5, the value yielding maximum robustness and the value of **k** which we set to q for all the algorithms. Our evaluation criterion is the empirical bias, a quantitative measure of robustness of a fit. In Appendix C we explain the motivation for

this choice (in the Online Resources we also report the results obtained using alternative evaluation criteria).

3.1 Empirical bias

Given an elliptical distribution \mathcal{E}_p with location vector $\boldsymbol{\mu}^u$ and covariance matrix $\boldsymbol{\Sigma}^u$ (the superscript u stands for uncontaminated) and an arbitrary distribution \mathcal{F}^c (the superscript c stands for contaminated), consider the ε -contaminated model

$$\mathcal{F}^\varepsilon = (1 - \varepsilon)\mathcal{E}_p(\boldsymbol{\mu}^u, \boldsymbol{\Sigma}^u) + \varepsilon\mathcal{F}^c.$$

For a fixed $q < p$ denote $\boldsymbol{\Sigma}_q^u$ the rank q approximation of $\boldsymbol{\Sigma}^u$ and $\mathbf{V}_q = \mathbf{P}_q \mathbf{L}_q \mathbf{P}_q^\top$ an estimator of $\boldsymbol{\Sigma}_q^u$. The (empirical) bias measures the difference between \mathbf{V}_q and $\boldsymbol{\Sigma}_q^u$. For this, we will consider more specifically the shape component of this difference which is called the shape bias. Given these two (rank reduced) covariance matrices, recall that the corresponding shape matrices are defined by $\boldsymbol{\Gamma}^u = |\boldsymbol{\Sigma}^u|^{-1/q} \boldsymbol{\Sigma}_q^u$ and $\mathbf{G}_q = |\mathbf{V}_q|^{-1/q} \mathbf{V}_q$. For an estimator of \mathbf{V}_q , all the information about the shape bias is contained in the matrix $(\boldsymbol{\Gamma}^u)^{-1/2} \mathbf{G}_q (\boldsymbol{\Gamma}^u)^{-1/2}$, or equivalently its condition number (Yohai and Maronna, 1990):

$$\text{bias}(\mathbf{V}_q) = \log \frac{\lambda_1((\boldsymbol{\Gamma}^u)^{-1/2} \mathbf{G}_q (\boldsymbol{\Gamma}^u)^{-1/2})}{\lambda_q((\boldsymbol{\Gamma}^u)^{-1/2} \mathbf{G}_q (\boldsymbol{\Gamma}^u)^{-1/2})},$$

where $\lambda_1(\lambda_q)$ is the largest (q^{th}) eigenvalue of the positive-semidefinite matrix $(\boldsymbol{\Gamma}^u)^{-1/2} \mathbf{G}_q (\boldsymbol{\Gamma}^u)^{-1/2}$. Evaluating the maximum bias of \mathbf{V}_q is an empirical matter: for a given sample, it depends on the dimensionality of the data, the rate of contamination by outliers, the distance separating them from the uncontaminated observations, and the spatial configuration of the outliers (\mathcal{F}^c). However, because all the algorithms we compare are rotation and shift equivariant, w.l.o.g. we can focus on configurations where $\boldsymbol{\Sigma}^u$ is diagonal, and $\boldsymbol{\mu}^u = \mathbf{0}_p$ (a p -vector of zeros), greatly reducing the number of scenarios we need to consider. Since the effect of contamination is presumably most harmful when the outlier belongs to the subspace spanned by $\boldsymbol{\Pi}_q^{u\perp}$ (the orthogonal complement of $\boldsymbol{\Pi}_q^u$) we, concentrate on the class of outlier configurations satisfying these conditions (Maronna, 2005). In the simulation results shown in Section 3.3, the outliers belong to the subspace spanned by the eigenvector corresponding to the $(q + 1)$ -th eigenvalue of $\boldsymbol{\Sigma}^u$ (as in Hubert et al. (2005)) whereas the simulation settings shown in the Online Resources the outliers belong to the subspace spanned by all the components of $\boldsymbol{\Pi}_q^{u\perp}$ (as is done in Maronna (2005)).

3.2 Outlier configurations

To quantify the robustness of the four algorithms, we generate many contaminated data sets \mathbf{X}^ε of size n with $\mathbf{X}^\varepsilon = \mathbf{X}^u \cup \mathbf{X}^c$ where \mathbf{X}^u and \mathbf{X}^c are the uncontaminated and contaminated part of the sample. In all simulations, the center of the uncontaminated data $\boldsymbol{\mu}^u = \mathbf{0}_{1 \times p}$ its $\boldsymbol{\Sigma}^u$ is either $\boldsymbol{\Sigma}^u$ or $10^{-4}\boldsymbol{\Sigma}^u$. We show results where $p \in \{100, 400\}$, $q \in \{5, 10, 15\}$, $n = 200$, and ε is one of $\{0.1, 0.2, 0.3, 0.4\}$.

To facilitate comparison, we consider a generalization to arbitrary values of q of the parametrization of $\boldsymbol{\Sigma}^u$ used in (Hubert et al. , 2005). To define this matrix, \mathbf{L} , we set the values of the first q elements of the diagonal of $\boldsymbol{\Sigma}^u$ so that they decrease exponentially and do not drop abruptly before the remaining, smaller, entries. More precisely, the first q entries of the diagonal of $\boldsymbol{\Sigma}^u$ are the first q Fibonacci numbers and the entries $q + 1, \dots, p$ are linearly decreasing as $(0.1, \dots, 0.001)$. In Section 2 of the Online Resources, we also provide results using a covariance matrix proposed by (Maronna , 2005).

Our measure of robustness, the bias, depends on the distance between the outliers and the non outlying observations which we will measure by

$$\nu = \min_{i \in I^c} \sqrt{(\mathbf{x}_i^\top (\boldsymbol{\Sigma}^u)^{-1} \mathbf{x}_i) / \chi_{0.975, p}^2}, \quad (10)$$

where I^c is an indicator for the observations coming from \mathbf{X}^c . (A more detailed description of how we set the location of the outliers can be found in Section 3 of the Online Resources.) In the simulations, the distance ν separating the outliers from the good data is one of $\{1, \dots, 10\}$.

We consider two outlier configurations frequently used in the robust PCA literature (Hubert et al. , 2005; Maronna , 2005): (a) Shift outliers: $\boldsymbol{\Sigma}^c = \boldsymbol{\Sigma}^u$ and $\boldsymbol{\mu}^c$ chosen to satisfy Equation (10); (b) Point-mass outliers: all the outliers are concentrated around a single point at a distance ν from \mathbf{X}^u . To obtain this effect, we set $\boldsymbol{\Sigma}^c = 10^{-4}\boldsymbol{\Sigma}^u$. Both of these outlier configurations are relevant in practical applications where they are, for example, similar to certain types of sensor faults and contamination scenarios.

For FastHCS, the number of initial $(q + 1)$ -subsets M is given as in Equation (1), with $e/n = 0.6$. The `rrcov` implementations for ROBPCA and PcaPP include hardcoded values for the number of starting subsets presumed by their authors to be sufficient for these methods. PcaL does not require starting subsets. Section 4 of the Online Resources explains how the reader can use code we supply to replicate all results from this section.

In Figures 1 to 2, we display the bias curves as lattice plots (Deepayan , 2008) for discrete combinations of p , q and ε . In all cases, we expect the outlier detection problem to become monotonically harder as we increase q and ε , so little information will be lost by considering a discrete grid of a few values for these parameters. The configurations also depend on the distance separating the data from the outliers. Here, the effects of ν on the bias are harder to foresee: clearly nearby outliers will be harder to detect but misclassifying distant outliers will increase the bias more. Therefore, we will test the algorithms for many values (and chart the results as a function) of ν . For each algorithm, a solid colored line will depict the median, and a dotted line (of the same color) the 75th percentile of bias(\mathbf{V}_q). Each figure is based on 12000 simulations.

3.3 Simulation results

Figure 1 displays the bias curves corresponding to the fits found by the algorithms for $p = 100$ for the shift (right) and point-mass (left) configurations. Regardless of the spatial configuration of the outliers or the value of ε , the fits found by PcaPP and PcaL generally have high values of bias(\mathbf{V}_q). As it turns out, PcaPP and PcaL will show poor performance on all of the remaining simulations as well. Since this poor performance is consistent, we will not discuss it in detail. The performance of ROBPCA is substantially better than the previous two algorithms, but it becomes increasingly unreliable as q increases. FastHCS shows almost no bias. Furthermore, we see that in some cases even after the bias curves of ROBPCA have re-descended, they are still above those of FastHCS. Given that this gap increases with ε , we infer that the eigenvalue estimation of ROBPCA is still influenced by the outliers, even when the eigenvectors are correctly estimated. FastHCS estimates both correctly.

We next consider the high dimensional case of $p > n$. Figure 2. Across all scenarios, the results are comparable to those in seen in Figure 1. FastHCS is the best performing method, being unaffected by the outliers, while the other methods show high biases on most settings.

Over all of the scenarios, FastHCS shows almost no bias, despite challenging outlier configurations. Furthermore, the bias curves corresponding to the fits found by FastHCS are also less variable: throughout, the 75th percentile of the bias corresponding to the FastHCS fit is typically closer to the median bias than is the case for the other algorithms. These findings indicate that FastHCS meets Criteria (4)-(5) for a robust PCA method. In contrast, we find that the

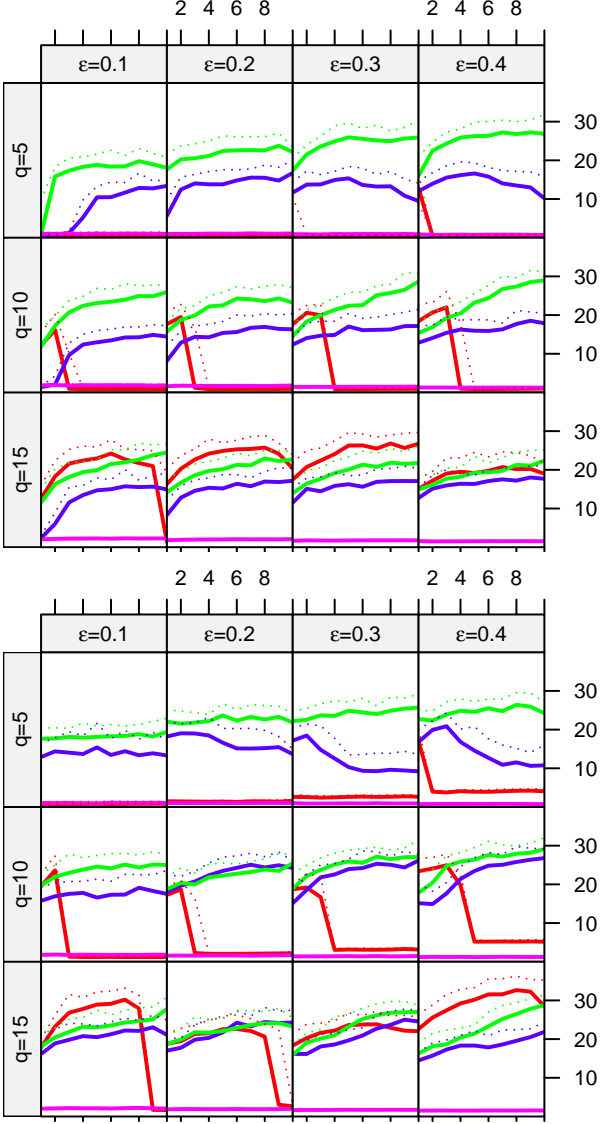


Fig. 1 $\text{bias}(\mathbf{V}_q)$ for $p = 100$, shift (top) and point-mass (bottom) as a function of ν . [ROBPCA](#), [PcaPP](#), [PcaL](#), [FastHCS](#).

performance of the other methods vary with the configuration of the outliers, the rate of the contamination, and the dimensionality of the q -subspace. In Section 5 of the Online Resources, we re-analyze these simulation results, giving similar plots for a measure of outlier misclassification, as well as the principal angle measure and $\text{bias}(\mathbf{P}_q)$, two measures of quality of fit focusing on the loadings.

An extended simulation study shows that the results we present above are robust the choice of different simulation settings (for example, those used in (Maronna, 2005)), and different bias measurements criterions. However, since the outcome of the extended study is nearly identical to the one we present in this section, we have relegated these results in the Online Resources.

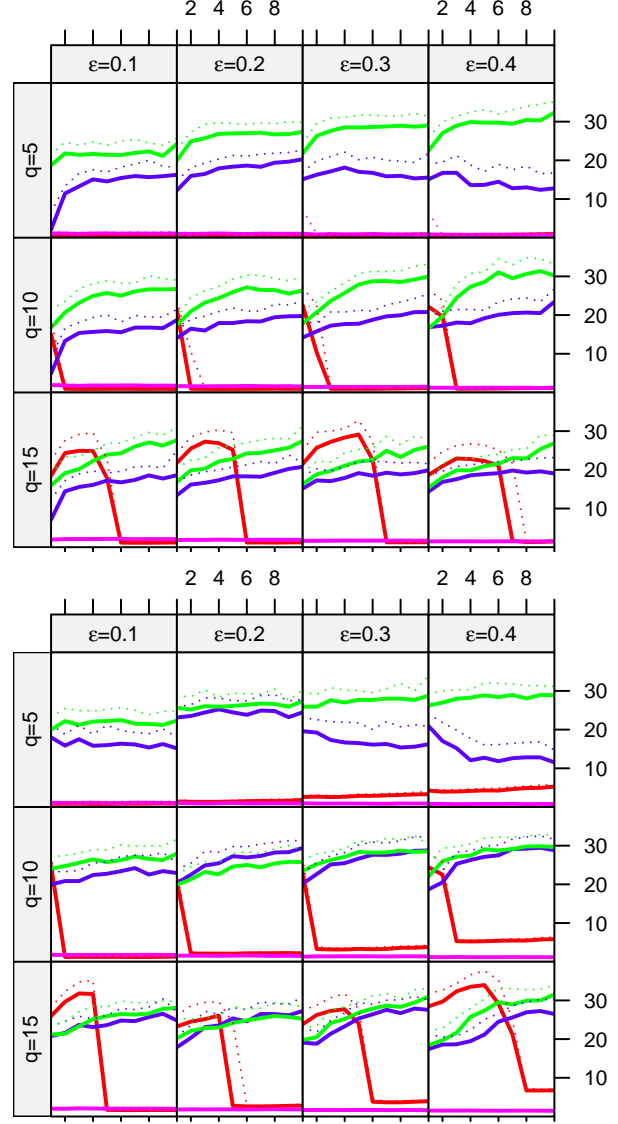


Fig. 2 $\text{bias}(\mathbf{V}_q)$ for $p = 400$, shift (top) and point-mass (bottom) as a function of ν . [ROBPCA](#), [PcaPP](#), [PcaL](#), [FastHCS](#).

4 Real data applications

We next apply the algorithms to three real data examples. We selected these examples because in each the observations in the data can be separated into two subgroups from which we construct a majority and an outlier group. They are taken from three fields that regularly use PCA: character recognition, chemometrics and genetics. A feature shared by all of these data sets is that the variables within each are measured on the same scale. Data sets with this property were selected to remove the ancillary problem of data standardization. If the variables are not on the same scale, it is common practice in PCA modelling to standardize the data, but the choice of how to do so robustly adds a

layer of subjectivity to the results. For the interested reader, the data sets used in this section are included in the **FastHCS** package. Section 6 of the Online Resources explains how the reader can use code we provide to replicate all results in this section.

The implementations of ROBPCA and PcaPP we use do not have an option to set the seed, but to ensure reproducibility of the results for FastHCS, we set `seed=1`. PcaL is a deterministic algorithm and uses no seeds. As in the simulations, we run each algorithm with default settings, except for the `alpha` parameter in ROBPCA which we set to 0.5. To illustrate the outlier detection capabilities of the algorithms, we display diagnostics plots. These show the OD and SD values for each observation, divided by the cut-off values in Equations (8) and (9) to put each of the methods on a comparable scale.

4.1 Selecting the number of components

We recommend using as large a value of q as possible for FastHCS, since this improves its outlier detection performance. However, to avoid the curse of dimensionality, it is also advised to set $q < n/5$ (Hubert et al. , 2005). In all the examples that follow, we select a relatively high number of components, $q = 15$, to strike a balance between computation time and accuracy. Once the outliers have been detected, components with large eigenvalues can be analyzed and used to construct a PCA model of the good data. One may also wish to use a selection criterion, such as the scree chart or contribution to variance. In Section 7 of the Online Resources, we also show results using the latter of these approaches in an extended analysis. In that analysis, the chemometrics data set illustrates how robust PCA methods parametrized based on a parsimonious, eigenvalue-based criterion, may miss outliers on minor components, even when the majority of the data may be modelled using a parsimonious model.

4.2 The Multiple Features data set

The Multiple Features data set (Van Breukelen et al. , 1998) contains many replications of hand written numerals ('0'-'9') extracted from nine original maps of a Dutch public utility. For each numeral, we have 200 replications (the observations) expressed as a vector of 76 of Fourier coefficients (the features) describing its shape. Finally, each numeral has been manually identified, yielding an extra vector of class labels. In this application, we will combine the vectors of Fourier coefficients corresponding to the 200 replications of the

digit '1' to the vector of Fourier coefficients corresponding to the first 150 replications of the digit '0' (so that $n = 350$ and $p = 76$). The goal of the methods will be to distinguish the '0's and the '1's.

To give an impression of the differences between the two groups, we plot the Fourier coefficients corresponding to the main (outlier) subgroup in the bottom (top) panels of Figure 3 as dark blue (light orange) curves. In general, the curves corresponding to the members of the two groups are visually similar. In particular, the vertical ranges of both largely overlap, and both sets of curves exhibit a similar pattern of variance clustering where the central 40 Fourier coefficients have systematically less dispersion than higher or lower ones.

Figure 4 depicts the four resulting diagnostic plots. We assign to each observation a color (dark blue or light orange) and a plot symbol (round or triangle) depending on whether the corresponding curve describes a member of class '1' or '0', respectively. The outlier plots of PcaPP and PcaL show that neither method makes any distinction between the two digits, and observations from both groups influence the corresponding PCA models. ROBPCA discovers a different structure in the data, confounding the '0's as the majority group and the '1's as outliers on the model space. Since only a few '1's are OD outliers, almost all of the observations influence the fitted loadings. In contrast to the other methods, FastHCS correctly identifies all of the '0's as outliers and identifies some '1's that might warrant additional scrutiny.

4.3 The Tablet data set

The Tablet data (Dyrby et al. , 2002) contains the results of an analysis on Escitalopram[®] tablets from the pharmaceutical company H. Lundbeck A/S using near-infrared (NIR) spectroscopy. The study includes tablets of four different dosages from pilot, laboratory and full scale production settings are included. Each tablet (the observations) is measured along 404 wavelengths (the variables). From this data, we extract two subsets of observations which we combine to obtain a new data set formed of two heterogeneous subgroups. Tablets of 80mg will make up the majority group and the rows corresponding to the first 50 tablets with a nominal weight of 250mg will serve as the outliers. This gives a high-dimensional data set (i.e. $p > n$) with $n = 130$, $p = 404$ and a contamination rate of $\varepsilon = 38\%$.

Figure 5, depicts the spectra of the 250mg (light orange) and 80mg (dark blue) tablets. The spectra of the 250mg tablets follow a different multivariate pattern than those of the 80mg tablets. For example, the spectra of the former are generally lower and more spread

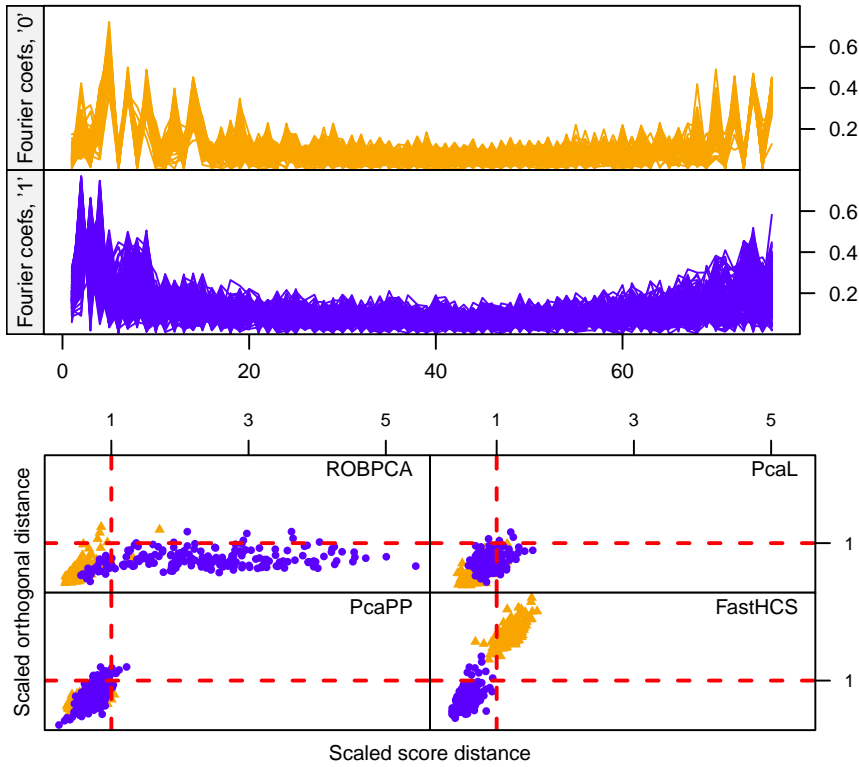


Fig. 3 The 350 vectors of Fourier coefficients of the character shapes for the Multiple Features data set. The first 150 curves (corresponding to observations with labels '0') are shown in the top panel in light orange. The Main group (200 curves) corresponding to observations with labels '1' are shown in the bottom panel.

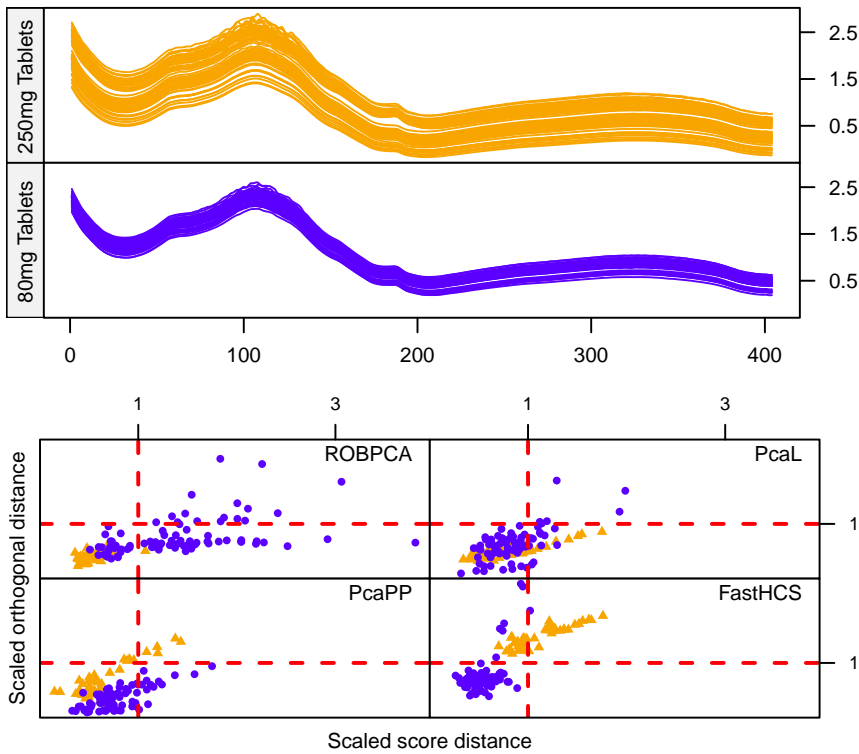


Fig. 5 Spectra of fifty 250mg (light orange) tablets and eighty 80mg (dark blue) tablets for the Tablet data set.

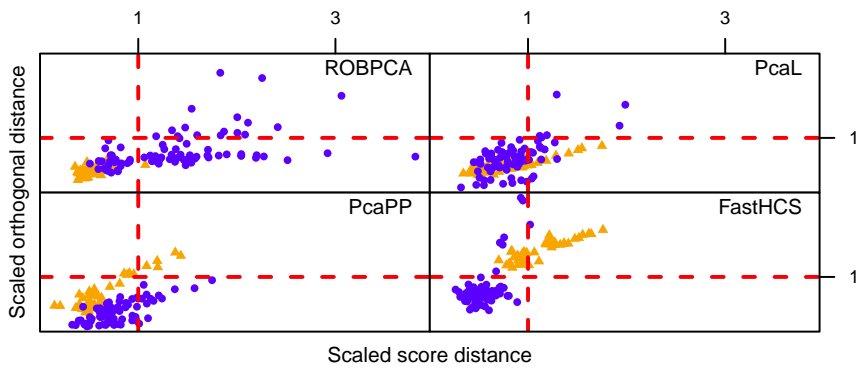


Fig. 6 Diagnostic plots of the scaled score and orthogonal distances of the measured spectra corresponding to the four robust PCA fits for the Tablet data set. The observations corresponding to 80mg (250mg) tablets are shown as dark blue circles (light orange triangles).

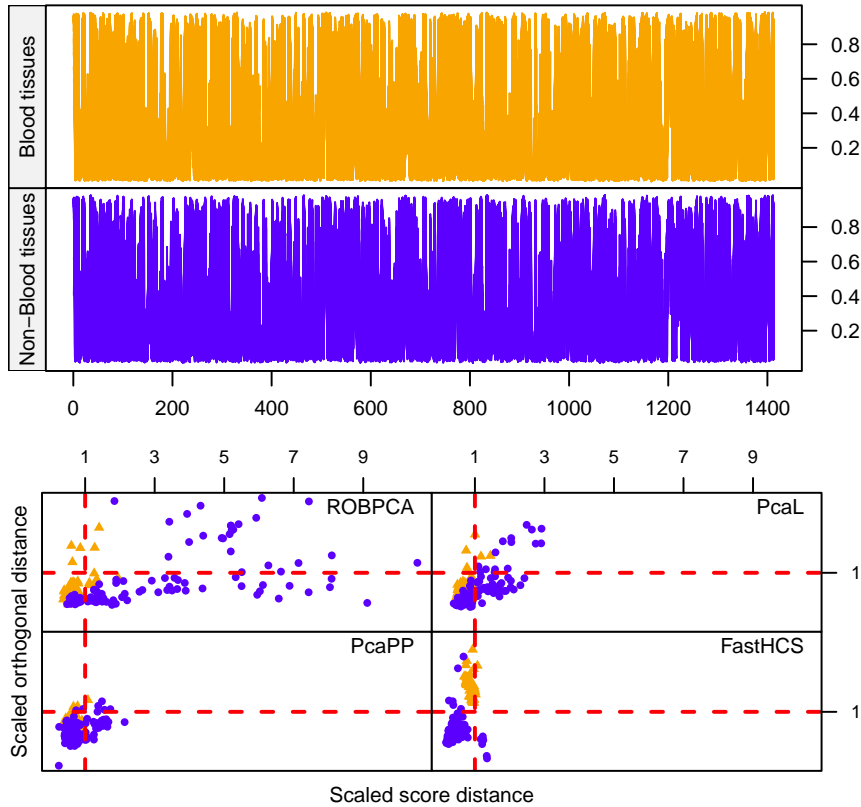


Fig. 7 The 198 vectors of cytosine methylation β values for the DNA alteration data set. The first 85 curves (corresponding to observations taken from blood tissues) are shown in the top panel in light orange. The main group (113 curves) corresponding to observations taken from non-blood, tissues are shown in the bottom panel in dark blue.

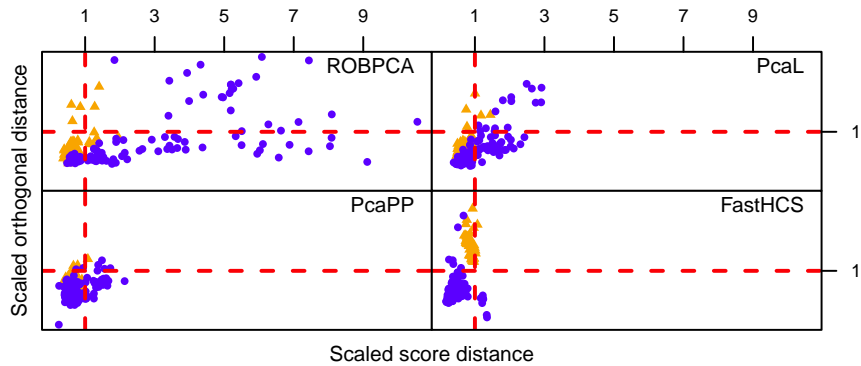


Fig. 8 Diagnostic plots of the scaled score and orthogonal distances of cytosine methylation β values corresponding to the four robust PCA algorithms for the DNA alteration data set. The observations with labels "non-blood" ("blood") are shown as dark blue circles (light orange triangles).

out than the spectra of the latter. Dyrby et al. (2002) explain that accurate models for NIR analyses of medical tablets are valuable for quality control purposes, since they are fast, nondestructive, noninvasive, and require little preparation. The goal of the algorithms will be to fit a model to the 80mg tablets, despite the presence in the sample of many 250mg tablets.

Figure 6 depicts the diagnostic plots of the scaled outlyingness measures obtained from each of the algorithms. To enhance the distinction between the two groups in our data, we show the 80mg tablets as (dark) blue circles and the 250mg tablets as (light) orange triangles. These results are similar to those we saw when we examined the Multiple Features data set. Again PcaPP and PcaL do not distinguish between the two groups and ROBPCA uses both groups to fit the loadings parameters and confuses the outliers with the majority group on the model space. In contrast, the diagnostic plot derived from the FastHCS fit establishes that the 250mg tablets do not follow the same multivariate patterns as 80mg tablets and, in fact, depart significantly from it. In the plot, we see that FastHCS assigns the outliers high OD values; excluding them from loadings and eigenvalue estimation. It also assigns many of them high SD values; revealing their distance on the model space.

4.4 DNA Alteration data set

In our final case study, we examine another high-dimensional data set; the DNA Alteration data set (Christensen et al., 2009). This data set consists of cytosine methylation β values collected at 1413 autosomal CpG loci (the variables) in a sample of 217 non-pathological human tissue specimens (the observations) taken from 10 different anatosites. In Christensen et al. (2009), the authors show that the tissue samples in this data set form three well separated subgroups. The first of these constitutes all 113 observations corresponding to cytosine methylation β values measured on "non-blood, non-placenta" (henceforth, simply "non-blood") tissues. A second subgroup of data points comprises the 85 cytosine methylation β measurements taken on blood tissues.

In this application, we will combine the 113 measurements of cytosine methylation β values corresponding to the samples "non-blood" tissue with 85 measurements taken from blood tissues (so that $n = 198$ and $p = 1413$). In Figure 7, we plot the 1413 β values corresponding to each blood (light orange) and non-blood (dark blue) observation. Visually, the curves of these two groups appear difficult to distinguish from one another. In particular, the vertical range of both overlap

and the groups do not exhibit any pronounced difference in the variability of the variables.

The diagnostic plot for PcaPP (Figure 8) reveals that the fitted model regards blood and non-blood tissue to be quantitatively similar. ROBPCA and PcaL also detect almost none of the outliers, but additionally consider a number of the non-blood observations to be SD outliers. As in the previous case studies, FastHCS correctly identifies all of the outliers. As a consequence, the parameter estimates corresponding to this model are more likely to reflect the true structure of non-blood tissue than those fitted by the other algorithms.

5 Outlook

In this article we introduced FastHCS to satisfy a number of criteria we expect a robust PCA method to have. In both the simulations and real data examples we performed, FastHCS met all of these criteria. In contrast, state-of-the-art methods did not, and often produced results one would expect from a non-robust method. This may seem like an extreme outcome, but it is in fact the very nature of dealing with outliers: if a method fails to identify them, the resulting model fit is often profoundly changed.

It is interesting to compare the performance of FastHCS and ROBPCA because these methods both use variants of projection pursuit. While FastHCS compares the fit produced by the I -index to that from the projection pursuit criterion, ROBPCA relies completely on the projection pursuit criterion to construct its initial subset. Thus, the difference in performance between FastHCS and ROBPCA that we observe in our simulations and real data examples arises from the fact that FastHCS nearly always chooses the I -index subset over the projection pursuit one.

In most applications, admittedly, data settings and contamination patterns will not be as difficult as those we featured in our simulations and real data examples, and in these easier cases the different methods will, hopefully, concur. Nevertheless, in three real data examples from fields where PCA is widely used, we were able to establish that real world situations can be challenging enough to push current state-of-the-art outlier detection procedures to their limits and beyond, justifying the development of better solutions. In any case, given that in practice we do not know the configuration of the outliers, as data analysts, we prefer to carry out our inferences while planning for the worst contingencies.

6 Acknowledgements

The authors wish to acknowledge the helpful comments from two anonymous referees and the editor which improved this paper.

A Vulnerability of the I -index to orthogonal outliers

Throughout this appendix, let \mathbf{Y} be an $n \times p$ data matrix of uncontaminated observations drawn from a rank q distribution \mathcal{F} , with q and integer satisfying $2 < q < \min(p, n)$. However, we do not observe \mathbf{Y} but an $n \times p$ (potentially) corrupted data matrix \mathbf{Y}^ε that consists of $g < n$ observations from \mathbf{Y} and $c = n - g$ arbitrary values with $\varepsilon = c/n$ denoting the (unknown) rate of contamination. Throughout, $h = \lceil (n + q + 1)/2 \rceil$ and the PCA estimates $(\mathbf{t}^I, \mathbf{L}_q^I, \mathbf{P}_q^I)$ are defined as in Section (2) with $(\mathbf{L}_q^I)_j, 1 \leq j \leq q$ will denoting the j -th diagonal entry of \mathbf{L}_q^I .

We will consider the finite sample breakdown (Donoho, 1982) in the context of PCA following (Li and Chen, 1985):

$$\varepsilon_1 = \min_{1 \leq c \leq n} \{ \varepsilon = \frac{c}{n} : (\mathbf{L}_q)_1 = \infty \} \quad (11)$$

$$\varepsilon_2 = \min_{1 \leq c \leq n} \{ \varepsilon = \frac{c}{n} : (\mathbf{L}_q)_q = 0 \} \quad (12)$$

Equation (11) defines the so-called finite sample explosion breakdown point and Equation (12) the so-called finite sample implosion breakdown point of PCA estimates $(\mathbf{t}, \mathbf{L}_q, \mathbf{P}_q)$, and the general finite sample breakdown point is $\varepsilon_n^* = \min(\varepsilon_1, \varepsilon_2)$.

The following assumptions (as per, for example Tyler (1994)) all pertain to the original, uncontaminated, data set \mathbf{Y} . We will consider the case whereby the point cloud formed by \mathbf{Y} lies in *general position* in \mathbb{R}^q . The following definition of *general position* is adapted from Rousseeuw and Leroy (1987):

DEFINITION 1: *General position in \mathbb{R}^q .* \mathbf{Y} is in general position in \mathbb{R}^q if no more than q -points of \mathbf{Y} lie in any $(q - 1)$ -dimensional affine subspace. For q -dimensional data, this means that there are no more than q points of \mathbf{Y} on any hyperplane, so that any $q + 1$ points of \mathbf{Y} always determine a q -simplex with non-zero determinant.

The I -index is shift invariant so that, w.l.o.g., we only consider cases where the good observations are centered at the origin. Throughout, we will also assume that the members of \mathbf{Y} are bounded:

$$\max_{i=1}^n \|\mathbf{y}_i\| < U_0$$

for some bounded scalar U_0 depending only on the uncontaminated observations and that the uncontaminated observations contain no duplicates:

$$\|\mathbf{y}_i - \mathbf{y}_j\| > 0 \quad \forall 1 \leq i < j \leq n.$$

A.1 Theorem 1: The implosion breakdown,

$$\varepsilon_2(\mathbf{t}^I, \mathbf{L}_q^I, \mathbf{P}_q^I), \text{ is } (n - h + 1)/n$$

Proof If at least h rows of \mathbf{Y}^ε are in general position in \mathbb{R}^q , any subset of h observations will contain at least $q + 1$ observations in general position. This guarantees that the q^{th}

eigenvalue corresponding to any h -subset is non-zero (Seber, 2008). Thus, it follows that $\varepsilon_2(\mathbf{t}^I, \mathbf{L}_q^I, \mathbf{P}_q^I) = (n - h + 1)/n$.

A.2 Finite sample explosion breakdown of $(\mathbf{t}^I, \mathbf{L}_q^I, \mathbf{P}_q^I)$

Denote $\mathbf{z} \in \mathbb{R}^p$ the outlying entries of \mathbf{Y}^ε and $\mathbf{z}^m = \|\mathbf{zP}_0^m\|$. The only outliers capable of causing explosion breakdown must satisfy:

$$\|\mathbf{z}\| \geq U_1, \quad (13)$$

$$\min_m \|\mathbf{zP}_0^m\| \leq U_2. \quad (14)$$

for any bounded scalar U_1 and U_2 depending only on the uncontaminated observations.

Proof Suppose that the outliers do not satisfy Equation (13) so that $\max_i \|\mathbf{y}_i^\varepsilon\| \leq U_1$, but that the PCA estimates $(\mathbf{t}^I, \mathbf{L}_q^I, \mathbf{P}_q^I)$ break down. This leads to a contradiction since

$$(\mathbf{L}_q^I)_1 \leq \max_{i \in H^I} \|\mathbf{y}_i^\varepsilon\| \quad (15)$$

Therefore, for a contaminated h -subset to cause explosion breakdown, the outliers must satisfy Equation (13).

Assume that an outlier \mathbf{z} does not satisfy Condition (14). Schmitt et al. (2014) showed that any h subset H^m indexing \mathbf{z} will have an unbounded value of $I(H^m, \mathbf{S}_0^m)$ if and only if \mathbf{z}^m is unbounded. But for the uncontaminated data, it holds that

$$\max_i \min_m \|\mathbf{y}_i \mathbf{P}_0^m\| \leq U_2 \quad (16)$$

so if the contaminated data set \mathbf{Y}^ε contains at least h entries from the original data matrix \mathbf{Y} , then it is always possible to construct a subset H^m of entries of \mathbf{Y}^ε for which $I(H^m, \mathbf{S}_0^m)$ is bounded so that H^m will never be selected over H^I .

B The finite sample breakdown point of FastHCS

In this appendix, we derive the finite sample breakdown point of FastHCS. Define \mathbf{Y} , \mathbf{Y}^ε and ε_n^* as in Appendix A. Recall that

$$D(\mathbf{Y}^\varepsilon, H^I, H^{PP}) = \frac{q}{j=1} \log \frac{\text{ave}_{i \in H^I} ((\mathbf{y}_i^\varepsilon - \mathbf{t}^I) \mathbf{P}_j^I)^2}{\text{var}_{i \in H^\bullet} ((\mathbf{y}_i^\varepsilon - \mathbf{t}^I) \mathbf{P}_j^I)^2} - \max_{j=1}^q \log \frac{\text{ave}_{i \in H^\bullet} ((\mathbf{y}_i^\varepsilon - \mathbf{t}^{PP}) \mathbf{P}_j^{PP})^2}{\text{var}_{i \in H^-} ((\mathbf{y}_i^\varepsilon - \mathbf{t}^{PP}) \mathbf{P}_j^{PP})^2}, \quad (17)$$

where $H^- = H^{PP} \setminus H^I$. Then, if $D(\mathbf{Y}^\varepsilon, H^I, H^{PP}) > 0$ or if $\max_{j=1}^q \text{var}_{i \in H^-} ((\mathbf{y}_i^\varepsilon - \mathbf{t}^{PP}) \mathbf{P}_j^{PP}) = 0$ then the final FastHCS estimates are based on H^{PP} . Otherwise, they are based on H^I .

Lemma 1 If $\|\mathbf{y}_i^\varepsilon\| > U_1$ and $\varepsilon < (n - 1)/2n$, then $i \notin H^\bullet$.

Proof (Debruyne and Hubert, 2009) showed that the population breakdown point of $(\mathbf{t}^{PP}, \mathbf{L}_q^{PP}, \mathbf{P}_q^{PP})$ is 50%, which corresponds to a finite sample breakdown point of $(n - 1)/2n$. Consequently, H^{PP} will not index any data point for which $\|\mathbf{y}_i^\varepsilon\| > U_1$. Since H^\bullet indexes the overlap between H^I and H^{PP} , if $\|\mathbf{y}_i^\varepsilon\| > U_1$, then $i \notin H^\bullet$.

Lemma 2 When \mathbf{Y} is in general position, $n > q > 2$, and $\varepsilon < \varepsilon_1 = (n - 1)/2n$, $(\mathbf{L}_q^I)_1 < \infty$.

Proof We will proceed by showing that the denominators in Equation (17) are bounded, while only the numerator dependent on H^{PP} is bounded.

Lemma 1 implies there exists a fixed constant U_4 such that

$$\|\mathbf{y}_i^\varepsilon \mathbf{P}_j\| < U_4 \quad \forall i \in H^\bullet, 1 \leq j \leq q \quad (18)$$

for any orthogonal matrix \mathbf{P} . Similarly, since the projection pursuit approach has a breakdown point of $(n - 1)/2n$, there exists a fixed U_5 such that

$$\|\mathbf{y}_i^\varepsilon \mathbf{P}_j\| < U_5 \quad \forall i \in H^{PP}, 1 \leq j \leq q \quad (19)$$

As a consequence of (18) and (19), there exists a fixed constant U_6 such that:

$$\sum_j \log(\text{var}_{i \in H^\bullet} ((\mathbf{y}_i^\varepsilon - \mathbf{t}^I) \mathbf{P}_j^I)^2) < U_6 \quad (20)$$

$$\sum_j \log(\text{var}_{i \in H^-} ((\mathbf{y}_i^\varepsilon - \mathbf{t}^{PP}) \mathbf{P}_j^{PP})^2) < U_6.$$

Next, note that

$$\max_j \log(\text{ave}_{i \in H^I} ((\mathbf{y}_i^\varepsilon - \mathbf{t}^I) \mathbf{P}_j^I)^2) = (\mathbf{L}_q^I)_1 \quad (21)$$

$$\min_j \log(\text{ave}_{i \in H^I} ((\mathbf{y}_i^\varepsilon - \mathbf{t}^I) \mathbf{P}_j^I)^2) = (\mathbf{L}_q^I)_q \geq \epsilon > 0, \quad (22)$$

(Equation (22) follows from Appendix A, Theorem 1), so that

$$\sum_j \log(\text{ave}_{i \in H^I} ((\mathbf{y}_i^\varepsilon - \mathbf{t}^I) \mathbf{P}_j^I)^2) \quad (23)$$

is not bounded from above. Conversely, $(\mathbf{t}^{PP}, \mathbf{L}_q^{PP}, \mathbf{P}_q^{PP})$ has an explosion breakdown point of $(n - 1)/2n$, so that there exists a fixed U_8 such that:

$$\sum_j \log(\text{ave}_{i \in H^\bullet} ((\mathbf{y}_i^\varepsilon - \mathbf{t}^{PP}) \mathbf{P}_j^{PP})^2) < U_8. \quad (24)$$

From Equations (20) and the unboundedness of (23) it follows that the left-hand side in Equation (17) is unbounded. However, by Equations (20) and (24), the right-hand side of Equation (17) is bounded from above so that in cases where outliers cause explosion breakdown of $(\mathbf{t}^I, \mathbf{L}_q^I, \mathbf{P}_q^I)$, criterion (17) will select $H^* = H^{PP}$. Since the breakdown point of $(\mathbf{t}^{PP}, \mathbf{L}_q^{PP}, \mathbf{P}_q^{PP})$ is $(n - 1)/2n$, we have that $\varepsilon_1 = (n - 1)/2n$.

Lemma 3 When \mathbf{Y} is in general position, $n > q > 2$, and $\varepsilon < \varepsilon_2 = (n - h + 1)/n$, then $(\mathbf{L}_q^I)_q > 0$.

Proof By Appendix A, Theorem 1, we have that the implosion breakdown point of $(\mathbf{t}^I, \mathbf{L}_q^I, \mathbf{P}_q^I)$ is $(n - h + 1)/n$. The implosion breakdown point of $(\mathbf{t}^{PP}, \mathbf{L}_q^{PP}, \mathbf{P}_q^{PP})$ is $(n - 1)/2n$, which is higher, so it follows that $\varepsilon_2 = (n - h + 1)/n$.

Theorem 1 For $n > p + 1 > 2$, the finite sample breakdown point of \mathbf{L}_q is

$$\varepsilon_n^*(\mathbf{L}_q, \mathbf{Y}^\varepsilon) = (n - h + 1)/n.$$

Proof The finite sample breakdown point of $\mathbf{L}_q = \min(\varepsilon_1, \varepsilon_2)$. Given Lemmas 2 and 3, $\min((n - 1)/2n, (n - h + 1)/n) = (n - h + 1)/n$.

C Measures of dissimilarity for robust PCA fits.

The objective of the simulation studies in Section 3.3 is to measure how much the fitted PCA parameters $(\mathbf{t}, \mathbf{L}_q, \mathbf{P}_q)$ obtained by four robust PCA methods deviate from the true $(\boldsymbol{\mu}^u, \mathbf{A}_q^u, \boldsymbol{\Pi}_q^u)$ when they are exposed to outliers. One way to compare PCA fits is with respect to their eigenvectors, as in the *maxsub* criterion (Björck and Golub, 1973):

$$\text{maxsub}(\mathbf{P}_q) = \arccos(\lambda_q^{1/2}(\mathbf{D}_q)),$$

where $\lambda_q(\mathbf{D}_q)$ is the smallest eigenvalue of the matrix $\mathbf{D}_q = \boldsymbol{\Pi}_q^\top \mathbf{P}_q \mathbf{P}_q^\top \boldsymbol{\Pi}_q$. The *maxsub* has an appealing geometrical interpretation as it represents the maximum angle between a vector in $\boldsymbol{\Pi}_q$ and the vector most parallel to it in \mathbf{P}_q . However, it does not exhaustively account for the dissimilarity between two sets of eigenvectors. As an alternative to the *maxsub*, Krzanowski (1979) proposes the total dissimilarity:

$$\text{sumsub}(\mathbf{P}_q) = \sum_{j=1}^q \lambda_j(\mathbf{D}_q), \quad (25)$$

which is an exhaustive measure of dissimilarity for orthogonal matrices. Furthermore, because $\sum_{j=1}^q \lambda_j(\mathbf{D}_q) = \text{Tr}(\mathbf{D}_q)$ and $|\mathbf{D}_q| = 1$ (Krzanowski, 1979), it is readily seen that (25) is a measure of sphericity of \mathbf{D}_q (it is proportional to the likelihood ratio test statistics for non-sphericity of \mathbf{D}_q (Muirhead, 1982, p. 333-335)). However, note that (25) now forfeits the geometric interpretation enjoyed by the *maxsub*.

In any case, measures of dissimilarity based solely on eigenvectors, such as the *maxsub* or *sumsub*, necessarily fail to account for bias in the estimation of the eigenvalues. This is problematic when used to evaluate robust fits because it is possible for outliers to exert substantially more influence on \mathbf{L}_q than on \mathbf{P}_q . An extreme example is given by the so-called good leverage type of contamination in which the outliers lie on the subspace spanned by $\boldsymbol{\Pi}_q$ so that even the classical PCA estimate (whose eigenvalues can be made arbitrarily bad by such outliers) will have low values of *maxsub*(\mathbf{P}_q).

In contrast, we are interested in an exhaustive measure of dissimilarity; one that summarizes the effects of the outliers on all the parameters of the PCA fit into a single number, so that the algorithms can be ranked in terms total dissimilarity. To construct such a measure, it is logical to base it on $\boldsymbol{\Sigma}_q^u = \boldsymbol{\Pi}_q^u \mathbf{A}_q^u (\mathbf{P}_q^u)^\top$ and its estimate $\mathbf{V}_q = \mathbf{P}_q \mathbf{L}_q \mathbf{P}_q^\top$ because they contain all the parameters of the fitted model. For our purposes, one need to only consider the effects of outliers on $\mathbf{G}_q = |\mathbf{V}_q|^{-1/q} \mathbf{V}_q$, the shape component of \mathbf{V}_q (Hubert et al., 2014). This is because to rank the observations in a contaminated sample in terms of their true outlyingness (and thus reveal the outliers), it is sufficient to estimate the shape component of $\boldsymbol{\Sigma}_q^u$ correctly. Consequently, an exhaustive measure of dissimilarity between \mathbf{G}_q and $\boldsymbol{\Gamma}_q = |\boldsymbol{\Sigma}_q^u|^{-1/q} \boldsymbol{\Sigma}_q^u$ is given by $\phi((\boldsymbol{\Gamma}_q^u)^{-1/2} \mathbf{G}_q (\boldsymbol{\Gamma}_q^u)^{-1/2})$, where ϕ is any measure of non-sphericity of its argument. In practice several choices of ϕ are possible, the simplest being the condition number of \mathbf{W} which is defined as the ratio of the largest to the smallest eigenvalue of \mathbf{W} (Maronna and Yohai, 1995), explaining the definition of *bias*(\mathbf{V}_q).

References

- Björck, Å. and Golub, G. H. (1973). Numerical Methods for Computing Angles Between Linear Subspaces. *Mathematics of Computation*, 27, 2, 579–594.
- Christensen, B.C Houseman, E.A. Marsit, C.J. Zheng, S. Wrench, M.R. Wiemels, J.L. Nelson, H.H. Karagas, M.R. Padbury, J.F. Bueno, R. Sugarbaker, D.J Yeh, R., Wiencke, J.K. Kelsey, K.T. (2009). Aging and Environmental Exposure Alter Tissue-Specific DNA Methylation Dependent upon CpG Island Context. *PLoS Genetics* 5(8), e1000602.
- Croux, C. and Ruiz-Gazen, A. (2005). High breakdown estimators for principal components: the projection-pursuit approach revisited. *Journal of Multivariate Analysis*, 95, 206–226.
- Donoho, D.L. (1982). Breakdown properties of multivariate location estimators. Ph.D. Qualifying Paper Harvard University.
- Debruyne, M. and Hubert, M. (2009). The influence function of the Stahel-Donoho covariance estimator of smallest outlyingness. *Statistics & probability letters* 79(3), 275–282.
- Deepayan, S. (2008). *Lattice: Multivariate Data Visualization with R*. Springer, New York.
- Dyrby, M. Engelsen, S.B. Nørgaard, L. Bruhn, M. and Lundsberg Nielsen, L. (2002). Chemometric Quantitation of the Active Substance in a Pharmaceutical Tablet Using Near Infrared (NIR) Transmittance and NIR FT Raman Spectra *Applied Spectroscopy* 56(5): 579–585.
- Hubert, M. Rousseeuw, P. J. and Vanden Branden, K. (2005). ROBPCA: a new approach to robust principal components analysis. *Technometrics*, 47, 64–79.
- Hubert, M., Rousseeuw, P. and Vakili, K. (2014). Shape bias of robust covariance estimators: an empirical study. *Statistical Papers*, Volume 55, Issue 1, pp 15–28.
- Jensen, D. R. (1986), The Structure of Ellipsoidal Distributions, II. Principal Components. *Biometrical Journal*, 28: 363–369.
- Jolliffe, I.T. (2002). *Principal Component Analysis*. Springer, New York. Second Edition.
- Krzanowski, W.J. (1979). Between-Groups Comparison of Principal Components. *Journal of the American Statistical Association*, Vol. 74, No. 367, pp. 703–707.
- Li, G., Chen, Z. (1985). Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and Monte Carlo. *Journal of the American Statistical Association*, 80, pp. 759–766.
- Locantore, N., Marron, J. S., Simpson, D. G., Tripoli, N., Zhang, J. T., and Cohen, K. L. (1999). Robust principal component analysis for functional data.

- Test. 8(1), 1–73.
- Maronna R. A. and Yohai V.J. (1995). The Behavior of the Stahel-Donoho Robust Multivariate Estimator. *Journal of the American Statistical Association* 90 (429), 330–341.
- Maronna, R. (2005). Principal Components and Orthogonal Regression Based on Robust Scales. *Technometrics*, 47, 264–273.
- Maronna, R. A.; Martin, R. D. and Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. Wiley, New York.
- Muirhead, R.J. (1982). *Aspects of Multivariate Statistical Theory*. John Wiley and Sons, New York.
- R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.
- Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. Wiley, New York.
- Schmitt, E. Öllerer, V. and Vakili, K. (2014). The finite sample breakdown point of PCS. *Statistics & Probability Letters*, 94, 214–220.
- Seber, G. A. F. (2008). *Matrix Handbook for Statisticians*. Wiley Series in Probability and Statistics. Wiley, New York.
- Stahel W. (1981). Breakdown of Covariance Estimators. Research Report 31, Fachgrupp für Statistik, E.T.H. Zürich.
- Todorov V. and Filzmoser P. (2009). An Object-Oriented Framework for Robust Multivariate Analysis. *Journal of Statistical Software*, 32, 1–47.
- Tyler, D.E. (1994). Finite Sample Breakdown Points of Projection Based Multivariate Location and Scatter Statistics.
- Vakili, K. and Schmitt, E. (2014). Finding multivariate outliers with FastPCS. *Computational Statistics & Data Analysis*, Vol. 69, 54–66.
- Van Breukelen, M. Duin, R.P.W. Tax, D.M.J. and Den Hartog, J.E. (1998). Handwritten digit recognition by combined classifiers. *Kybernetika*, 34, 381–386.
- Wu, W., Massart, D. L., and de Jong, S. (1997), The Kernel PCA Algorithms for Wide Data. Part I: Theory and Algorithms. *Chemometrics and Intelligent Laboratory Systems*, 36, 165–172.
- Yohai, V.J. and Maronna, R.A. (1990). The Maximum Bias of Robust Covariances. *Communications in Statistics—Theory and Methods*, 19, 2925–2933.